

Statistical significance in high-dimensional linear models

PETER BÜHLMANN^{1,*}

¹*Seminar für Statistik, HG G17, ETH Zürich, CH-8092 Zürich, Switzerland*

E-mail: *buhlmann@stat.math.ethz.ch

February 7, 2012

We propose a method for constructing p-values for general hypotheses in a high-dimensional linear model. The hypotheses can be local for testing a single regression parameter or they may be more global involving several up to all parameters. Furthermore, when considering many hypotheses, we show how to adjust for multiple testing taking dependence among the p-values into account. Our technique is based on Ridge estimation with an additional correction term due to a substantial projection bias in high dimensions. We prove strong error control for our p-values and provide sufficient conditions for detection: for the former, we do not make any assumption on the size of the true underlying regression coefficients. We demonstrate the method in simulated examples and a real data application.

Keywords: Global testing, Lasso, Multiple testing, Ridge regression, Variable selection, Westfall-Young permutation procedure.

1. Introduction

Many data problems nowadays carry the structure that the number p of covariables may greatly exceed sample size n , i.e., $p \gg n$. In such a setting, a huge amount of work has been pursued addressing prediction of a new response variable, estimation of an underlying parameter-vector and variable selection, see for example the books by [Hastie et al. \(2009\)](#), [Bühlmann and van de Geer \(2011\)](#) or the more specific review article by [Fan and Lv \(2010\)](#). With a few exceptions, see Section 1.3.1, the proposed methods and presented mathematical theory do not address the problem of assigning uncertainties, statistical significance or confidence: thus, the part of statistical hypothesis testing and construction of confidence intervals is largely unexplored and underdeveloped. Yet, such significance or confidence measure are crucial in applications where interpretation of parameters and variables is very important. The focus of this paper is the construction of p-values and corresponding multiple testing adjustment for a high-dimensional linear model which is often very useful in $p \gg n$ settings:

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon, \tag{1.1}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, \mathbf{X} is a fixed design $n \times p$ design matrix, β^0 is the true underlying $p \times 1$ parameter vector and ε is the $n \times 1$ stochastic error vector with $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. having $\mathbb{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2 < \infty$; throughout the paper, p may be much larger n .

We are interested in testing one or many null-hypotheses of the form:

$$H_{0,G}; \beta_j^0 = 0 \text{ for all } j \in G, \quad (1.2)$$

where $G \subseteq \{1, \dots, p\}$ is a subset of all the indices of the covariables. Of substantial interest is the case where $G = \{j\}$ corresponding to a hypothesis for the individual j th regression parameter ($j = 1, \dots, p$). At the other end of the spectrum is the global null-hypothesis where $G = \{1, \dots, p\}$, and we allow for any G between an individual and the global hypothesis.

1.1. Past work about high-dimensional linear models

We review in this section an important stream of research for high-dimensional linear models. The more familiar reader may skip Section 1.1.

1.1.1. The Lasso

The Lasso (Tibshirani, 1996)

$$\hat{\beta}_{\text{Lasso}} = \hat{\beta}_{\text{Lasso}}(\lambda) = \arg\min_{\beta} (\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1),$$

has become tremendously popular for estimation in high-dimensional linear models. The three main themes which have been considered in the past are prediction of the regression surface (and for a new response variable) with corresponding measure of accuracy

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}} - \beta^0)\|_2^2/n, \quad (1.3)$$

estimation of the parameter vector whose quality is judged by

$$\|\hat{\beta}_{\text{Lasso}} - \beta^0\|_q \quad (q \in \{1, 2\}), \quad (1.4)$$

and variable selection or estimating the support of β^0 , denoted by the active set $S_0 = \{j; \beta_j^0 \neq 0, j = 1, \dots, p\}$ such that

$$\mathbb{P}[\hat{S} = S_0] \quad (1.5)$$

is large for a selection (estimation) procedure \hat{S} .

Greenshtein and Ritov (2004) proved the first result closely related to prediction as measured in (1.3). Without any conditions on the deterministic design matrix \mathbf{X} one has with high probability at least $1 - 2\exp(-t^2/2)$:

$$\begin{aligned} \|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n &\leq 3/2\lambda\|\beta^0\|_1, \\ \lambda &= 4\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}, \end{aligned} \quad (1.6)$$

see [Bühlmann and van de Geer \(2011, Cor.6.1\)](#). Thereby, we assume Gaussian errors but such an assumption can be relaxed ([Bühlmann and van de Geer, 2011](#), formula (6.5)). From an asymptotic point of view (where p and n diverge to ∞), the regularization parameter $\lambda \asymp \sqrt{\log(p)/n}$ leads to consistency for prediction if the truth is sparse with respect to the ℓ_1 -norm such that $\|\beta^0\|_1 = o(\lambda^{-1}) = o(\sqrt{n/\log(p)})$. The convergence rate is then at best $O_P(\lambda) = O_P(\sqrt{\log(p)/n})$.

Such a slow rate of convergence can be improved under additional assumptions on the design matrix \mathbf{X} . The ill-posedness of the design matrix can be quantified using the concept of “modified” eigenvalues. Consider the matrix $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$. The smallest eigenvalue of $\hat{\Sigma}$ is

$$\lambda_{\min}(\hat{\Sigma}) = \min_{\beta} \beta^T \hat{\Sigma} \beta.$$

Of course, $\lambda_{\min}(\hat{\Sigma})$ equals zero if $p > n$. Instead of taking the minimum on the right-hand-side over all $p \times 1$ vectors β , we replace it by a *constrained* minimum, typically over a cone. This leads to the concept of restricted eigenvalues ([Bickel et al., 2009](#); [Koltchinskii, 2009a,b](#); [Raskutti et al., 2010](#)) or weaker forms such as the compatibility constants ([van de Geer, 2007](#)) or further slight weakening of the latter ([Sun and Zhang, 2011](#)). Relations among the different conditions and “modified” eigenvalues are discussed in [van de Geer and Bühlmann \(2009\)](#) and [Bühlmann and van de Geer \(2011, Ch.6.13\)](#). Assuming that the smallest “modified” eigenvalue is larger than zero, one can derive an oracle inequality of the following prototype: with probability at least $1 - 2\exp(-t^2/2)$ and using λ as in (1.6):

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n + \lambda\|\hat{\beta}_{\text{Lasso}} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2, \quad (1.7)$$

where ϕ_0 is the compatibility constant (smallest “modified” eigenvalue) of the fixed design matrix \mathbf{X} ([Bühlmann and van de Geer, 2011, Cor.6.2](#)). Again, this holds by assuming Gaussian errors but the result can be extended to non-Gaussian distributions. From (1.7), we have two immediate implications: from an asymptotic point of view, using $\lambda \asymp \sqrt{\log(p)/n}$ and assuming that ϕ_0 is bounded away from 0,

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n = O_P(s_0 \log(p)/n), \quad (1.8)$$

$$\|\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0\|_1 = O_P(s_0 \sqrt{\log(p)/n}), \quad (1.9)$$

i.e., a fast convergence rate for prediction as in (1.8) and an ℓ_1 -norm bound for the estimation error. We note that the oracle convergence rate, where an oracle would know the active set S_0 , is $O_P(s_0/n)$: the $\log(p)$ -factor is the price to pay by not knowing the active set S_0 . An ℓ_2 -norm bound can be derived as well: $\|\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0\|_2 = O_P(\sqrt{s_0 \log(p)/n})$ assuming a slightly stronger restricted eigenvalue condition. Results along these lines have been established by [Bunea et al. \(2007\)](#), [van de Geer \(2008\)](#) who covers generalized linear models as well, [Zhang and Huang \(2008\)](#), [Meinshausen and Yu \(2009\)](#), [Bickel et al. \(2009\)](#) among others.

The Lasso is doing variable selection: a simple estimator of the active set S_0 is $\hat{S}_{\text{Lasso}}(\lambda) = \{j; \hat{\beta}_{\text{Lasso};j}(\lambda) \neq 0\}$. In order that $\hat{S}_{\text{Lasso}}(\lambda)$ has a good accuracy for S_0 , we have to require that the non-zero regression coefficients are sufficiently large (since otherwise, we cannot detect the variables in S_0 with high probability). We make a “beta-min” assumption whose asymptotic form reads as

$$\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}. \quad (1.10)$$

Furthermore, when making a restrictive assumption for the design, called neighborhood stability, or assuming the equivalent irrepresentable condition, and choosing a suitable $\lambda \gg \sqrt{\log(p)/n}$:

$$\mathbb{P}[\hat{S}_{\text{Lasso}}(\lambda) = S_0] \rightarrow 1,$$

see [Meinshausen and Bühlmann \(2006\)](#), [Zhao and Yu \(2006\)](#), and [Wainwright \(2009\)](#) establishes exact scaling results. The “beta-min” assumption in (1.10) as well as the irrepresentable condition on the design are restrictive and non-checkable. Furthermore, these conditions are essentially necessary ([Meinshausen and Bühlmann, 2006](#); [Zhao and Yu, 2006](#)). Thus, under weaker assumptions, we can only derive a weaker yet useful result about variable screening. Assuming a restricted eigenvalue condition on the fixed design \mathbf{X} and assuming the “beta-min” condition in (1.10) we still have asymptotically that for $\lambda \asymp \sqrt{\log(p)/n}$:

$$\mathbb{P}[\hat{S}(\lambda) \subseteq S_0] \rightarrow 1 \quad (n \rightarrow \infty). \quad (1.11)$$

The cardinality of the estimated active set (typically) satisfies $|\hat{S}(\lambda)| \leq \min(n, p)$: thus if $p \gg n$, we achieve a massive and often useful dimensionality reduction in the original covariates.

We summarize that a slow convergence rate for prediction “always” holds. Assuming some “constrained minimal eigenvalue” condition on the fixed design \mathbf{X} , we obtain the fast convergence rate in (1.8), and an estimation error bound as in (1.9); with the additional “beta-min” assumption, we obtain the practically useful variable screening property in (1.11). For consistent variable selection, we necessarily need a (much) stronger condition on the fixed design, and such a strong condition is questionable to be true in a practical problem. Hence variable selection might be a too ambitious goal with the Lasso. That is why the original translation of Lasso (Least Absolute Shrinkage and Selection Operator) may be better re-translated as Least Absolute Shrinkage and *Screening* Operator. We refer to [Bühlmann and van de Geer \(2011\)](#) for an extensive treatment of the properties of the Lasso.

1.1.2. Other methods

Of course, the three main inference tasks in a high-dimensional linear model, as described by (1.3), (1.4) and (1.5), can be pursued with other methods than the Lasso.

An interesting line of proposals include concave penalty functions instead of the ℓ_1 -norm in the Lasso ([Fan and Li, 2001](#)), and further developed by [Zhang \(2010\)](#). The

adaptive Lasso (Zou, 2006), analyzed in the high-dimensional setting by Huang et al. (2008) and van de Geer et al. (2011), can be interpreted as an approximation of some concave penalization approach (Zou and Li, 2008). A related procedure to the adaptive Lasso is the relaxed Lasso (Meinshausen, 2007). Another method is the Dantzig selector (Candès and Tao, 2007) which has similar statistical properties as the Lasso (Bickel et al., 2009). Other algorithms include orthogonal matching pursuit (which is essentially forward variable selection) or L_2 Boosting (matching pursuit) which have desirable properties (Tropp, 2004; Bühlmann, 2006).

Quite different from estimation of the high-dimensional parameter vector are variable screening procedures which aim for an analogous property as in (1.11). Prominent examples include the SIS-method (Fan and Lv, 2008) and high-dimensional variable screening or selection properties have been established for forward variable selection (Wang, 2009) and for the PC-algorithm (Bühlmann et al., 2010).

1.2. Assigning uncertainties and p-values for high-dimensional regression

At the core of statistical inference is the specification of statistical uncertainties, significance and confidence. For example, instead of having a variable selection result where the probability in (1.5) is large, we would like to have measures controlling a type I error (false positive selections), including p-values which are adjusted for large-scale multiple testing, or construction of confidence intervals or regions. In the high-dimensional setting, answers to these core goals are challenging.

Meinshausen and Bühlmann (2010) propose Stability Selection, a very generic method which is able to control the expected number of false positive selections: that is, denoting by $V = |\hat{S} \cap S_0^c|$, Stability Selection yields a finite-sample upper bound of $\mathbb{E}[V]$ (not only for linear models but for many other inference problems). To achieve this, a very restrictive (but presumably non-necessary) exchangeability condition is made which, in a linear model, is implied by a restrictive assumption for the design matrix. On the positive side, there is no requirement of a “beta-min” condition as in (1.10) and the method seems to give reliable control of $\mathbb{E}[V]$.

Wasserman and Roeder (2009) propose a procedure for variable selection based on sample splitting. Using their idea and extending it to multiple sample splitting, Meinshausen et al. (2009) develop a much more stable method for construction of p-values for hypotheses $H_{0,j} : \beta_j^0 = 0$ ($j = 1, \dots, p$) and for adjusting them in a non-naive way for multiple testing over p (dependent) tests. The main drawback of this procedure is its required “beta-min” assumption in (1.10). And this is very undesirable since for statistical hypothesis testing, the test should control type I error regardless of the size of the coefficients, while the power of the test should be large if the absolute value of the coefficient would be large: thus, we should avoid assuming (1.10).

Up to now, for the high-dimensional linear model case with $p \gg n$, it seems that only Zhang and Zhang (2011) managed to construct a procedure which leads to statistical tests for $H_{0,j}$, and even to confidence intervals for the true underlying parameters β_j^0 ,

without assuming a “beta-min” condition.

1.3. A loose description of our new results

We start with considering Ridge regression for estimating the high-dimensional regression parameter. We then develop a bias correction, addressing the issue that Ridge regression is estimating the regression coefficient vector projected to the row space of the design matrix: the corrected estimator is denoted by $\hat{\beta}_{\text{corr}}$.

Theorem 1 describes that under the null-hypothesis, the distribution of a suitably normalized $a_{n,p}|\hat{\beta}_{\text{corr}}|$ can be stochastically (componentwise) upper-bounded:

$$\begin{aligned} a_{n,p}|\hat{\beta}_{\text{corr}}| &\stackrel{\text{as}}{\preceq} (|Z_j| + \Delta_j)_{j=1}^p, \\ (Z_1, \dots, Z_p) &\sim \mathcal{N}_p(0, \sigma^2 n^{-1} \Omega), \end{aligned} \quad (1.12)$$

for some *known* positive definite matrix Ω and some *known* constants Δ_j . This is the key to derive p-values based on this stochastic upper bound. It can be used for construction of p-values for individual hypotheses $H_{0,j}$ as well as for more global hypotheses $H_{0,G}$ for *any* subset $G \subseteq \{1, \dots, p\}$, including cases where G is (very) large. Furthermore, Theorem 2 justifies a simple approach for controlling the familywise error rate when considering multiple testing of regression hypotheses. Our multiple testing adjustment method itself is closely related to the Westfall-Young permutation procedure (Westfall and Young, 1993) and hence, it offers high power, especially in presence of dependence among the many test-statistics (Meinshausen et al., 2011).

1.3.1. Relation to other work

Our new method as well as the approach in Zhang and Zhang (2011) provide p-values (and the latter also confidence intervals) without assuming a “beta-min” condition. Both of them build on using linear estimators and a correction using a non-linear initial estimator such as the Lasso. Using e.g. the Lasso directly leads to the problem of characterizing the distribution of the estimator (in a tractable form): this seems very difficult in high-dimensional settings while it has been worked out for low-dimensional problems (Knight and Fu, 2000). The work by Zhang and Zhang (2011) is the only one which studies (sufficiently closely) related questions and goals.

The approach by Zhang and Zhang (2011) is based on the idea of projecting the high-dimensional parameter vector to low-dimensional components, as occurring naturally in the hypotheses $H_{0,j}$ about single components, and then proceeding with a linear estimator. This idea is pursued with the “efficient score function” approach from semiparametric statistics (Bickel et al., 1998). The difficulty in the high-dimensional setting is the construction of the score vector z_j from which one can derive a confidence interval for β_j^0 : Zhang and Zhang (2011) propose it as the residual vector from the Lasso when regressing $\mathbf{X}^{(j)}$ against all other variables $\mathbf{X}^{(\setminus j)}$ (where $\mathbf{X}^{(J)}$ denotes the columns of the design matrix corresponding to index set $J \subseteq \{1, \dots, p\}$). They then prove the asymptotic validity of confidence intervals for finite, sparse linear combinations of β^0 . The difference

to our work is primarily a rather different construction of the projection where we make use of Ridge estimation with a very simple choice of regularization. Furthermore, our approach and analysis allows for the construction of p-values of hypotheses $H_{0,G}$ where the cardinality of G may be large, and we develop a simple method for multiple testing adjustment to control the familywise error rate. On the other hand, we only focus on hypotheses testing and we do not address the issue of constructing confidence intervals.

2. Model, estimation and p-values

Consider one or many null-hypotheses as in (1.2). We are interested in constructing p-values for hypotheses $H_{0,G}$ without imposing a “beta-min” condition as in (1.10): the statistical test itself will distinguish whether a regression coefficient is small or not.

2.1. Identifiability

We consider model (1.1) with fixed design. Without making additional assumptions on the design matrix \mathbf{X} , there is a problem of identifiability. Clearly, if $p > n$ and hence $\text{rank}(\mathbf{X}) \leq n < p$, there are different parameter vectors θ such that $\mathbf{X}\beta^0 = \mathbf{X}\theta$. Thus, we cannot identify β^0 from the distribution of Y_1, \dots, Y_n (and fixed design \mathbf{X}).

Shao and Deng (2011) give a characterization of identifiability in a high-dimensional linear model (1.1) with fixed design. Following their approach, it is useful to consider the singular value decomposition

$$\begin{aligned} \mathbf{X} &= RSV^T, \\ R &n \times n \text{ matrix with } U^T U = I_n, \\ S &n \times n \text{ diagonal matrix with singular values } s_1, \dots, s_n, \\ V &p \times n \text{ matrix with } V^T V = I_n. \end{aligned}$$

Denote by $\mathcal{R}(\mathbf{X}) \subset \mathbb{R}^p$ the linear space generated by the n rows of \mathbf{X} . The projection of \mathbb{R}^p onto $\mathcal{R}(\mathbf{X})$ is then

$$P_{\mathbf{X}} = \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^- \mathbf{X} = VV^T,$$

where A^- denotes the pseudo-inverse of a squared matrix A .

A natural choice of a parameter θ^0 such that $\mathbf{X}\beta^0 = \mathbf{X}\theta^0$ is the projection of β^0 onto $\mathcal{R}(\mathbf{X})$. Thus,

$$\theta^0 = P_{\mathbf{X}}\beta^0 = VV^T\beta^0. \quad (2.1)$$

Then, of course, $\beta^0 \in \mathcal{R}(\mathbf{X})$ if and only if $\beta^0 = \theta^0$.

2.2. Ridge regression

Consider Ridge regression

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_2^2 = (n^{-1}\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}n^{-1}\mathbf{X}^T\mathbf{Y}, \quad (2.2)$$

where $\lambda = \lambda_n$ is a regularization parameter. By construction of the estimator, $\hat{\beta} \in \mathcal{R}(\mathbf{X})$; and indeed, as discussed below, $\hat{\beta}$ is a reasonable estimator for $\theta^0 = P_X\beta^0$. We denote by

$$\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}.$$

The covariance matrix of the Ridge estimator, multiplied by n , is then

$$\Omega = \Omega(\lambda) = (\hat{\Sigma} + \lambda_n I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda_n I)^{-1},$$

a quantity which will appear at many places again. We assume that

$$\Omega_{\min}(\lambda) := \min_{j \in \{1, \dots, p\}} ((\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1})_{jj} > 0. \quad (2.3)$$

We do not require that $\Omega_{\min}(\lambda)$ is bounded away from zero as a function of n and p . Thus, the assumption in (2.3) is very mild: a rather peculiar design would be needed to violate the condition, see also the equivalent formulation in formula (2.4) below. Furthermore, (2.3) is easily checkable.

We denote by $\lambda_{\min \neq 0}(A)$ the smallest non-zero eigenvalue of a symmetric matrix A . We then have the following result.

Proposition 1. Consider the Ridge regression estimator $\hat{\beta}$ in (2.2) with regularization parameter $\lambda > 0$. Assume condition (2.3), see also (2.4). Then,

$$\begin{aligned} \max_{j \in \{1, \dots, p\}} |\mathbb{E}[\hat{\beta}_j] - \theta_j^0| &\leq \lambda \|\theta^0\|_2 \lambda_{\min \neq 0}(\hat{\Sigma})^{-1}, \\ \min_{j \in \{1, \dots, p\}} \operatorname{Var}(\hat{\beta}_j) &\geq n^{-1} \sigma^2 \Omega_{\min}(\lambda). \end{aligned}$$

A proof is given in Supplementary Section 8.1, relying in large parts on Shao and Deng (2011). We now discuss under which circumstances the estimation bias is smaller than the standard error. Qualitatively, this happens if $\lambda > 0$ is chosen sufficiently small. For a more quantitative discussion, we study the behavior of $\Omega_{\min}(\lambda)$ as a function of λ and we obtain an equivalent formulation of (2.3). We use the spectral decomposition of $\hat{\Sigma} = UDU^T$ with $UU^T = U^TU = I_p$ and $D = \operatorname{diag}(D_{11}, \dots, D_{pp})$ consisting of the ordered eigenvalues $D_{11} \leq D_{22} \leq \dots \leq D_{pp}$ of $\hat{\Sigma}$. Then,

$$\Omega = U \operatorname{diag}\left(\frac{D_{11}}{(D_{11} + \lambda)^2}, \dots, \frac{D_{pp}}{(D_{pp} + \lambda)^2}\right) U^T.$$

Denote by q the smallest index such that $D_{qq} > 0$, and thus $D_{qq} = \lambda_{\min \neq 0}(\hat{\Sigma})$: if $\operatorname{rank}(\mathbf{X}) = n < p$, then $q = p - n + 1$.

Lemma 1. We have the following:

1.

$$\Omega_{\min}(\lambda) = \min_j \sum_{r=q}^p \frac{D_{rr}}{(D_{rr} + \lambda)^2} U_{jr}^2.$$

From this we get:

$$(2.3) \text{ holds if and only if } \min_{1 \leq j \leq p} \max_{q \leq r \leq p} U_{jr}^2 > 0. \quad (2.4)$$

2. Assuming (2.3),

$$\Omega_{\min}(0^+) := \lim_{\lambda \searrow 0^+} \Omega_{\min}(\lambda) = \min_j \sum_{r=q}^p \frac{1}{D_{rr}} U_{jr}^2 > 0.$$

3.

$$\text{if (2.3) holds: } 0 < L_C \leq \liminf_{\lambda \in (0, C]} \Omega_{\min}(\lambda) \leq M_C < \infty, \quad (2.5)$$

for any $0 < C < \infty$, and where $0 < L_C < M_C < \infty$ are constants which depend on C (and on the design matrix \mathbf{X}).

The proof is straightforward using the spectral decomposition given above. From Proposition 1 we immediately obtain the following result.

Corollary 1. Consider the Ridge regression estimator $\hat{\beta}$ in (2.2) with regularization parameter $\lambda > 0$ satisfying

$$\lambda \Omega_{\min}(\lambda)^{-1/2} \leq n^{-1/2} \sigma \|\theta^0\|_2^{-1} \lambda_{\min \neq 0}(\hat{\Sigma}). \quad (2.6)$$

In addition, assume condition (2.3), see also (2.4). Then,

$$\max_{j \in \{1, \dots, p\}} (\mathbb{E}[\hat{\beta}_j] - \theta_j^0)^2 \leq \min_{j \in \{1, \dots, p\}} \text{Var}(\hat{\beta}_j).$$

Due to the third statement in Lemma 1 regarding the behavior of $\Omega_{\min}(\lambda)$, (2.6) can be fulfilled for a sufficiently small value of λ .

2.3. The projection bias and corrected Ridge regression

As discussed in Section 2.1, Ridge regression is estimating the parameter $\theta^0 = P_{\mathbf{X}} \beta^0$ given in (2.1). Thus, in general, besides the estimation bias governed by the choice of λ , there is an additional projection bias $B_j = \theta_j^0 - \beta_j^0$ ($j = 1, \dots, p$). Clearly,

$$B_j = (P_{\mathbf{X}} \beta^0)_j - \beta_j^0 = (P_{\mathbf{X}})_{jj} \beta_j^0 - \beta_j^0 + \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \beta_k^0.$$

In terms of constructing p-values, controlling type I error for testing $H_{0,j}$ or $H_{0,G}$ with $j \in G$, the projection bias has only a disturbing effect if $\beta_j^0 = 0$ and $\theta_j^0 \neq 0$, and we only have to consider the bias under the null-hypothesis:

$$B_{H_{0,j}} = \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \beta_k^0. \quad (2.7)$$

Under $H_{0,j}$, or $H_{0,G}$ with $j \in G$, we then have $B_{H_{0,j}} = \theta_j^0$. We can estimate $B_{H_{0,j}}$ by

$$\hat{B}_{H_{0,j}} = \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \hat{\beta}_{\text{init};k},$$

where $\hat{\beta}_{\text{init}}$ is an initial estimator such as the Lasso which guarantees a certain estimation accuracy, see assumption (A) below. This motivates the following bias-corrected Ridge estimator for testing $H_{0,j}$, or $H_{0,G}$ with $j \in G$:

$$\hat{\beta}_{\text{corr};j} = \hat{\beta}_j - \hat{B}_{H_{0,j}} = \hat{\beta}_j - \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \hat{\beta}_{\text{init};k}. \quad (2.8)$$

We then have the following representation.

Proposition 2. Assume model (1.1) with Gaussian errors. Consider the corrected Ridge regression estimator $\hat{\beta}_{\text{corr}}$ in (2.8) with regularization parameter $\lambda > 0$, and assume (2.3). Then,

$$\begin{aligned} \hat{\beta}_{\text{corr};j} &= Z_j + \gamma_j \quad (j = 1, \dots, p) \\ Z_1, \dots, Z_p &\sim \mathcal{N}_p(0, n^{-1} \sigma^2 \Omega), \quad \Omega = \Omega(\lambda), \\ \gamma_j &= (P_{\mathbf{X}})_{jj} \beta_j^0 - \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) + b_j(\lambda), \\ b_j(\lambda) &= \mathbb{E}[\hat{\beta}_j(\lambda)] - \theta_j^0. \end{aligned}$$

A proof is given in Supplementary Section 8.1. The normalizing factors for the variables Z_j bringing them to the $\mathcal{N}(0, 1)$ -scale are

$$a_{n,p;j}(\sigma) = n^{1/2} \sigma^{-1} \Omega_{jj}^{-1/2} \quad (j = 1, \dots, p)$$

which are also depending on λ through $\Omega = \Omega(\lambda)$. We refer to Section 4.1 where the unusually fast divergence of $a_{n,p;j}(\sigma)$ is discussed. The test-statistics we consider are simple functions of $a_{n,p;j}(\sigma) \hat{\beta}_{\text{corr};j}$.

2.4. Stochastic bound for the distribution of the corrected Ridge estimator: asymptotics

We provide here an asymptotic stochastic bound for the distribution of $a_{n,p;j}(\sigma) \hat{\beta}_{\text{corr};j}$ under the null-hypothesis. The asymptotic formulation is compact and the basis for the

construction of p-values in Section 2.5, but we give more detailed finite-sample results in Section 6.

We consider a triangular array of observations from a linear model as in (1.1):

$$\mathbf{Y}_n = \mathbf{X}_n \beta_n^0 + \varepsilon_n, \quad n = 1, 2, \dots, \quad (2.9)$$

where all the quantities and also the dimension $p = p_n$ are allowed to change with n . We make the following assumption.

(A) There are constants $\Delta_j = \Delta_{j,n} > 0$ such that

$$\mathbb{P}[\cap_{j=1}^{p_n} \{|a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0)| \leq \Delta_{j,n}\}] \rightarrow 1 \quad (n \rightarrow \infty).$$

We will discuss in Section 2.4.1 constructions for such bounds Δ_j . Our next result is the key to obtain a p-value for testing the null-hypothesis $H_{0,j}$ or $H_{0,G}$, saying that asymptotically,

$$a_{n,p;j}(\sigma) \hat{\beta}_{\text{corr};j} \stackrel{\text{as.}}{\preceq} |W| + \Delta_j,$$

where $W \sim \mathcal{N}(0, 1)$, and similarly for the multi-dimensional version with $\hat{\beta}_{\text{corr};G}$ (where \preceq denotes “stochastically smaller or equal to”).

Theorem 1. Assume model (2.9) with fixed design and Gaussian errors. Consider the corrected Ridge regression estimator $\hat{\beta}_{\text{corr}}$ in (2.8) with regularization parameter $\lambda_n > 0$ such that

$$\lambda_n \Omega_{\min}(\lambda_n)^{-1/2} = o(\min(n^{-1/2} \|\theta^0\|_2^{-1} \lambda_{\min \neq 0}(\hat{\Sigma}))) \quad (n \rightarrow \infty),$$

and assume condition (A) and (2.3) (while for the latter, the quantity does not need to be bounded away from zero). Then, for $j \in \{1, \dots, p_n\}$ and if $H_{0,j}$ holds: for all $u \in \mathbb{R}$,

$$\limsup_{n \rightarrow \infty} (\mathbb{P}[a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr};j}| > u] - \mathbb{P}[|W| + \Delta_j > u]) \leq 0,$$

where $W \sim \mathcal{N}(0, 1)$. Similarly, for any sequence of subsets $\{G_n\}_n$, $G_n \subseteq \{1, \dots, p_n\}$ and if H_{0,G_n} holds: for all $u \in \mathbb{R}^+$,

$$\limsup_{n \rightarrow \infty} (\mathbb{P}[\max_{j \in G_n} a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr};j}| > u] - \mathbb{P}[\max_{j \in G_n} (a_{n,p;j}(\sigma) |Z_j| + \Delta_j) > u]) \leq 0,$$

where Z_1, \dots, Z_P are as in Proposition 2.

A proof is given in Supplementary Section 8.1. We note that the distribution of $\max_{j \in G_n} (a_{n,p;j}(\sigma) |Z_j| + \Delta_j)$ does not depend on σ and can be easily computed via simulation.

2.4.1. Bounds Δ_j in assumption (A)

We discuss an approach for constructing the bounds Δ_j . As mentioned above, they should not involve any unknown quantities so that we can use them for constructing p-values from the distribution of $|W| + \Delta_j$ or $\max_{j \in G_n} (a_{n,p;j}(\sigma)|Z_j| + \Delta_j)$, respectively.

We rely on the (crude) bound

$$|a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0)| \leq a_{n,p;j}(\sigma) \max_{k \neq j} |(P_{\mathbf{X}})_{jk}| \|\hat{\beta}_{\text{init}} - \beta^0\|_1. \quad (2.10)$$

To proceed further, we consider the Lasso as initial estimator. Due to (1.7) we obtain

$$|a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0)| \leq \max_{k \neq j} |a_{n,p;j}(\sigma) (P_{\mathbf{X}})_{jk}| 4\lambda_{\text{Lasso}} s_0 \phi_0^{-2}, \quad (2.11)$$

where the last inequality holds on a set with probability at least $1 - 2\exp(-t^2/2)$ when choosing λ_{Lasso} as in (1.6). The assumptions we require are summarized next.

Lemma 2. Consider the linear model (2.9) with fixed design which satisfies the compatibility condition with constant $\phi_0^2 = \phi_{0,n}^2$. Consider the Lasso as initial estimator $\hat{\beta}_{\text{init}}$ with regularization parameter $\lambda_{\text{Lasso}} = 4\sigma\sqrt{C \log(p_n)/n}$ for some $2 < C < \infty$. Assume that the sparsity $s_0 = s_{0,n} = o((n/\log(p_n))^\xi)$ ($n \rightarrow \infty$) for some $0 < \xi < 1/2$, and that $\liminf_{n \rightarrow \infty} \phi_{0,n}^2 > 0$. Then,

$$\Delta_j := \max_{k \neq j} |a_{n,p;j}(\sigma) (P_{\mathbf{X}})_{jk}| (\log(p)/n)^{1/2-\xi} \quad (2.12)$$

satisfies assumption (A).

A proof follows from (2.11). When assuming bounded sparsity $s_{0,n} \leq M < \infty$ for all n , we can choose $\xi = 0$ with an additional constant M on the right-hand side of (2.12). In our practical examples, we use $\xi = 0.05$. We summarize the results as follows.

Corollary 2. Assume the conditions of Theorem 1 without condition (A) and the conditions of Lemma 2. Then, when using the Lasso as initial estimator, the statements in Theorem 1 hold.

2.5. P-values

Our construction of p-values is based on the asymptotic distributions in Theorem 1. For an individual hypothesis $H_{0,j}$, we define the p-value for the two-sided alternative as

$$P_j = 2(1 - \Phi((a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j}| - \Delta_j)_+)). \quad (2.13)$$

Of course, we could also consider one-sided alternatives with the obvious modification for P_j . For a more general hypothesis $H_{0,G}$ with $|G| > 1$, we use the maximum as test

statistics (but other statistics such as weighted sums could be chosen as well) and denote by

$$\begin{aligned}\hat{\gamma}_G &= \max_{j \in G} a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr},j}|, \\ J_G(c) &= \mathbb{P}[\max_{j \in G} (a_{n,p;j}(\sigma) |Z_j| + \Delta_j) \leq c],\end{aligned}$$

where the latter is independent of σ and can be easily computed via simulation (Z_1, \dots, Z_p are as in Proposition 2). Then, the p-value for $H_{0,G}$, against the alternative being the complement $H_{0,G}^c$, is defined as

$$P_G = 1 - J_G(\hat{\gamma}_G). \quad (2.14)$$

We note that when $\Delta_j \equiv \Delta$ is the same for all j , we can rewrite $P_G = 1 - \mathbb{P}[\max_{j \in G} a_{n,p;j}(\sigma) |Z_j| \leq (\hat{\gamma}_G - \Delta)_+]$ which is a direct analogue of (2.13).

Error control follows immediately by the construction of the p-values.

Corollary 3. Assume the conditions in Theorem 1. Then, for any $0 < \alpha < 1$,

$$\begin{aligned}\limsup_{n \rightarrow \infty} \mathbb{P}[P_j \leq \alpha] - \alpha &\leq 0 \text{ if } H_{0,j} \text{ holds,} \\ \limsup_{n \rightarrow \infty} \mathbb{P}[P_G \leq \alpha] - \alpha &\leq 0 \text{ if } H_{0,G} \text{ holds.}\end{aligned}$$

Furthermore, for any sequence $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) which converges sufficiently slowly, the statements also hold when replacing α by α_n .

A discussion about detection power of the method is given in Section 4. Further remarks about these p-values are given in Supplementary Section 8.4.

2.5.1. Estimation of σ

In practice, for the p-values in (2.13) and (2.14), we use the normalizing factor $a_{n,p;j}(\hat{\sigma})$ with an estimate $\hat{\sigma}$. These p-values are asymptotically controlling the type I error if $\mathbb{P}[\hat{\sigma} \geq \sigma] \rightarrow 1$ ($n \rightarrow \infty$). This follows immediately from the construction.

We propose to use the estimator $\hat{\sigma}$ from the Scaled Lasso method (Sun and Zhang, 2011). Assuming $s_0 \log(p)/n = o(1)$ ($n \rightarrow \infty$) and the compatibility condition for the design, Sun and Zhang (2011) prove that $|\hat{\sigma}/\sigma - 1| = o_P(1)$ ($n \rightarrow \infty$).

3. Multiple testing

We aim to strongly control the familywise error rate $\mathbb{P}[V > 0]$ where V is the number of false positive selections. For simplicity, we consider first individual hypotheses $H_{0,j}$ ($j \in \{1, \dots, p\}$). The generalization to multiple testing of general hypotheses $H_{0,G}$ with $|G| > 1$ is discussed in Section 3.2.

Based on the individual p-values P_j , we want to construct corrected p-values $P_{\text{corr};j}$ corresponding to the following decision rule:

$$\text{reject } H_{0,j} \text{ if } P_{\text{corr};j} \leq \alpha \ (0 < \alpha < 1).$$

We denote the associated estimated set of rejected hypotheses (the set of significant variables) by $\hat{S}_\alpha = \{j; P_{\text{corr};j} \leq \alpha\}$. Furthermore, recall that $S_0 = \{j; \beta_j^0 \neq 0\}$ is the set of true active variables. The number of false positives using the nominal significance level α is denoted by

$$V_\alpha = \hat{S}_\alpha \cap S_0^c.$$

The goal is to construct $P_{\text{corr};j}$ such that $\mathbb{P}[V_\alpha > 0] \leq \alpha$, or that the latter holds at least in an asymptotic sense. The method we describe here is closely related to the Westfall-Young procedure (Westfall and Young, 1993).

Consider the variables $Z_1, \dots, Z_p \sim \mathcal{N}_p(0, \sigma^2 n^{-1} \Omega)$ appearing in Proposition 2 or Theorem 1. Consider the following distribution function:

$$F_Z(c) = \mathbb{P}\left[\min_{1 \leq j \leq p} 2(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)) \leq c\right].$$

and define

$$P_{\text{corr};j} = F_Z(P_j + \zeta), \tag{3.1}$$

where $\zeta > 0$ is an arbitrarily small number, e.g. $\zeta = 0.01$ for using the method in practice. Regarding the choice of $\zeta = 0$ (which we use in all empirical examples in Section 5, see the Remark appearing after Theorem 2 below). The distribution function $F_Z(\cdot)$ is independent of σ and can be easily computed via simulation of the dependent, mean zero jointly Gaussian variables Z_1, \dots, Z_p . It is computationally (much) faster than simulation of the so-called minP-statistics (Westfall and Young, 1993) which would require fitting $\hat{\beta}_{\text{corr}}$ many times.

3.1. Asymptotic justification of the multiple testing procedure

We first derive familywise error control in an asymptotic sense. For a finite sample result, see Section 6. We consider the framework as in (2.9).

Theorem 2. Assume the conditions in Theorem 1. For the p-value in (2.13) and using the correction in (3.1) with $\zeta > 0$ we have: for $0 < \alpha < 1$,

$$\limsup_{n \rightarrow \infty} \mathbb{P}[V_\alpha > 0] \leq \alpha.$$

Furthermore, for any sequence $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) which converges sufficiently slowly, it holds that $\limsup_{n \rightarrow \infty} \mathbb{P}[V_{\alpha_n} > 0] - \alpha_n \leq 0$.

A proof is given in Supplementary Section 8.1.

Remark: Multiple testing correction in (3.1) with $\zeta = 0$. We could modify the correction in (3.1) using $\zeta = 0$: the statement in Theorem 2 can then be derived when making the additional assumption that

$$\sup_{n \in \mathbb{N}} \sup_u |G'_{n,Z}(u)| < \infty, \quad (3.2)$$

where $F_{n,Z}(\cdot) = F_Z(\cdot)$ is the distribution function appearing in (3.1) which depends in the asymptotic framework on n and (mainly on) $p = p_n$. Verifying (3.2) may not be easy for general matrices $\Omega = \Omega_{n,p_n}$. However, for the special case where Z_1, \dots, Z_p are independent,

$$G'_Z(u) = p\varphi(u)(1 - \Phi(u))^{p-1}$$

which is nicely bounded as a function of u , over all values of p .

3.2. Multiple testing of general hypotheses

The methodology for testing many general hypotheses H_{0,G_j} with $|G_j| \geq 1$, $j = 1, \dots, m$ is the same as before. Denote by $S_{0,G} = \{j; H_{0,G_j} \text{ does not hold}\}$ and by $S_{0,G}^c = \{j; H_{0,G_j} \text{ holds}\}$; note that these sets are determined by the true parameter vector β^0 . Since the p-value in (2.14) is of the form $P_{G_j} = 1 - J_{G_j}(\hat{\gamma}_{G_j})$, we consider

$$F_{G,Z} = \mathbb{P}[\min_{j=1,\dots,m} (1 - J_{G_j}(\gamma_{G_j,Z})) \leq c], \quad \gamma_{G,Z} = \max_{j \in G} (a_{n,p;j}(\sigma) |Z_j|)$$

which can be easily computed via simulation (and it is independent of σ). We then define the corrected p-value as

$$P_{\text{corr};G_j} = F_{G,Z}(P_{G_j} + \zeta),$$

where $\zeta > 0$ is a small value such as $\zeta = 0.01$; see also the definition in (3.1) and the corresponding discussion for the case where $\zeta = 0$ (which now applies to the distribution function $F_{G,Z}$ instead of F_Z). We denote by $\hat{S}_{G,\alpha} = \{j; P_{\text{corr};G_j} \leq \alpha\}$ and $V_{G,\alpha} = \hat{S}_{G,\alpha} \cap S_{0,G}^c$.

If $J_{G_j}(\cdot)$ has a bounded first derivative, for all j , we can obtain the same result, under the same conditions, as in Theorem 2 (and without making a condition on the cardinalities of G_j). If $J_{G_j}(\cdot)$ has not a bounded first derivative, we can get around this problem by modifying the p-value P_{G_j} in (2.14) to $\tilde{P}_{G_j} = 1 - J_{G_j}(\hat{\gamma}_{G_j} - \nu)$ for any (small) $\nu > 0$ and proceeding with \tilde{P}_{G_j} .

4. Sufficient conditions for detection

We consider detection of alternatives $H_{0,j}^c$ or $H_{0,G}^c$ with $|G| > 1$. We use again the notation S_0 as in Section 3 and denote by $a_n \gg b_n$ that $a_n/b_n \rightarrow \infty$ ($n \rightarrow \infty$).

Theorem 3. Consider the setting and assumptions as in Theorem 1.

1. When considering individual hypotheses $H_{0,j}$: if $j \in S_0$ with

$$|\beta_j^0| \gg a_{n,p;j}(\sigma)^{-1} |(P_{\mathbf{X}})_{jj}|^{-1} \max(\Delta_j, 1)$$

there exists an $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) such that

$$\mathbb{P}[P_j \leq \alpha_n] \rightarrow 1 \quad (n \rightarrow \infty),$$

while we still have for $j \in S_0^c$: $\limsup_{n \rightarrow \infty} \mathbb{P}[P_j \leq \alpha_n] - \alpha_n \leq 0$ (see Corollary 3).

2. When considering individual hypotheses $H_{0,G}$ with $G = G_n$ and $|G_n| > 1$: if $H_{0,G}^c$ holds, with

$$\max_{j \in G_n} |a_{n,p;j}(\sigma) P_{jj} \beta_j^0| \gg \max(\max_{j \in G_n} |\Delta_j|, \sqrt{\log(|G_n|)}),$$

there exists an $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) such that

$$\mathbb{P}[P_{G_n} \leq \alpha_n] \rightarrow 1 \quad (n \rightarrow \infty),$$

while if $H_{0,G}$ holds, $\limsup_{n \rightarrow \infty} \mathbb{P}[P_{G_n} \leq \alpha_n] - \alpha_n \leq 0$ (see Corollary 3).

3. When considering multiple hypotheses $H_{0,j}$: if for all $j \in S_0$,

$$|\beta_j^0| \gg a_{n,p;j}(\sigma)^{-1} |(P_{\mathbf{X}})_{jj}|^{-1} \max(\Delta_j, \sqrt{\log(p_n)})$$

there exists an $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) such that

$$\mathbb{P}[P_{\text{corr};j} \leq \alpha_n] \rightarrow 1 \quad (n \rightarrow \infty) \text{ for } j \in S_0$$

while we still have that $\limsup_{n \rightarrow \infty} \mathbb{P}[V_{\alpha_n} > 0] - \alpha_n \leq 0$ (see Theorem 2).

4. If in addition, $a_{n,p;j}(\sigma) \rightarrow \infty$ for all j appearing in the conditions on β_j^0 , we can replace in all the statements 1-3 the “ \gg ” relation by “ $\geq C$ ”, where $0 < C < \infty$ is a sufficiently large constant.

A proof is given in Supplementary Section 8.1. Under the additional assumption of Lemma 2, where the Lasso is used as initial estimator and using the bounds in (2.12), we obtain the bound (for statement 1 in Theorem 3):

$$|\beta_j^0| \geq C \max \left(\frac{\max_{k \neq j} |(P_{\mathbf{X}})_{jk}|}{|(P_{\mathbf{X}})_{jj}|} \left(\frac{\log(p_n)}{n} \right)^{1/2-\xi}, \frac{1}{|(P_{\mathbf{X}})_{jj}|} a_{n,p;j}(\sigma)^{-1} \right), \quad (4.1)$$

where $0 < \xi < 1/2$. This can be sharpened using the oracle bound, assuming known order of sparsity:

$$\Delta_{\text{orac};j} = D s_{0,n} \max_{k \neq j} a_{n,p;j}(\sigma) |(P_{\mathbf{X}})_{jk}| \sqrt{\log(p_n)/n}$$

for some $D > 0$ sufficiently large (for example, assuming $s_{0,n}$ is bounded, and replacing $s_{0,n}$ by 1 and choosing $D > 0$ sufficiently large). It then suffices to require

$$\begin{aligned} |\beta_j^0| &\geq C \max \left(\frac{\max_{k \neq j} |(P_{\mathbf{X}})_{jk}|}{|(P_{\mathbf{X}})_{jj}|} s_{0,n} \left(\frac{\log(p_n)}{n} \right)^{1/2}, \frac{1}{|(P_{\mathbf{X}})_{jj}| a_{n,p;j}(\sigma)} \right) \text{ for 1. in Th. 3,} \\ |\beta_j^0| &\geq C \max \left(\frac{\max_{k \neq j} |(P_{\mathbf{X}})_{jk}|}{|(P_{\mathbf{X}})_{jj}|} s_{0,n} \left(\frac{\log(p_n)}{n} \right)^{1/2}, \frac{\sqrt{\log(p_n)}}{|(P_{\mathbf{X}})_{jj}| a_{n,p;j}(\sigma)} \right) \text{ for 3. in Th. 3,} \end{aligned} \quad (4.2)$$

and analogously for the second statement in Theorem 3.

4.1. Order of magnitude of normalizing factors

The order of $a_{n,p;j}(\sigma)$ is typically much larger than \sqrt{n} since in high dimensions, Ω_{jj} is very small. This means that Ridge regression $\hat{\beta}_j$ has a much faster convergence rate than $1/\sqrt{n}$ for estimating the projected parameter θ_j^0 . This looks counter-intuitive at first sight: the reason for the phenomenon is that $\|\theta^0\|_2$ can be much smaller than $\|\beta^0\|_2$ and hence, Ridge regression (which estimates the parameter θ^0) is operating on a much smaller scale. This fact is essentially an implication of the first statement in Lemma 1 (without the “min_j” part) since the eigenvectors are normalized with $\sum_{r=1}^p U_{jr}^2 = 1$ and q is large when p is large. For further discussion about the fast convergence rate $a_{n,p;j}(\sigma)^{-1}$, see Supplementary Section 8.4.

While $a_{n,p;j}(\sigma)^{-1}$ is usually small, there is compensation with $(P_{\mathbf{X}})_{jj}^{-1}$ which can be rather large. In the detection bound in e.g. the first part of (4.2), both terms appearing in the maximum are often of the same order of magnitude; see also Figure 3 in Supplementary Section 8.4. Assuming such a balance of terms, we obtain in e.g. the first part of (4.2):

$$|\beta_j^0| \geq C \frac{\max_{k \neq j} |(P_{\mathbf{X}})_{jk}|}{|(P_{\mathbf{X}})_{jj}|} s_{0,n} \sqrt{\log(p_n)/n}.$$

The value of $\kappa_j = \max_{k \neq j} |(P_{\mathbf{X}})_{jk}|/|(P_{\mathbf{X}})_{jj}|$ is often a rather small number between 0.05 and 4, see Table 1 in Section 5. For comparison, Zhang and Zhang (2011) establish detection for single hypotheses $H_{0,j}$ with β_j^0 in the $1/\sqrt{n}$ range (our results in Theorem 3 also cover general hypotheses H_{0,G_n} with $|G_n|$ potentially large, as well as multiple testing). For the extreme case with $G_n = \{1, \dots, p_n\}$, we are in the setting of detection of the global hypotheses, see for example Ingster et al. (2010) for characterizing the detection boundary in case of independent covariables. Our analysis of detection is only providing sufficient conditions, for rather general (fixed) design matrices.

5. Numerical results

As initial estimator for $\hat{\beta}_{\text{corr}}$ in (2.8), we use the Scaled Lasso with scale-independent regularization parameter $\lambda_{\text{Lasso}} = 2\sqrt{\log(p)/n}$: it provides an initial estimate $\hat{\beta}_{\text{init}}$ as

well as an estimate $\hat{\sigma}$ for the standard deviation σ . The parameter λ for Ridge regression in (2.2) is always chosen as $\lambda = 1/n$, reflecting the assumption in Theorem 1 that it should be small.

For single testing, we construct p-values as in (2.13) or (2.14) with Δ_j from (2.12) with $\xi = 0.05$. For multiple testing with familywise error control, we consider p-values as in (3.1) with $\zeta = 0$ (and Δ_j as above).

5.1. Simulations

We simulate from the linear model as in (1.1) with $\varepsilon \sim \mathcal{N}_n(0, I)$, $n = 100$ and the following configurations:

- (M1) For both $p \in \{500, 2500\}$, the fixed design matrix is generated from a realization of n i.i.d. rows from $\mathcal{N}_p(0, I)$. Regarding the regression coefficients, we consider active sets $S_0 = \{1, 2, \dots, s_0\}$ with $s_0 \in \{3, 15\}$ and three different strengths of regression coefficients where $\beta_j^0 \equiv b$ ($j \in S_0$) with $b \in \{0.25, 0.5, 1\}$.
- (M2) The same as in (M1) but for both $p \in \{500, 2500\}$, the fixed design matrix is generated from a realization of n i.i.d. rows from $\mathcal{N}_p(0, \Sigma)$ with $\Sigma_{jk} \equiv 0.8$ ($j \neq k$) and $\Sigma_{jj} = 1$.

The resulting signal to noise ratios $\text{SNR} = \|\mathbf{X}\beta^0\|_2/\sigma$ are rather small:

$p \in \{500, 2500\}$	(3, 0.25)	(3, 0.5)	(3, 1)	(15, 0.25)	(15, 0.5)	(15, 1)
(M1)	0.46	0.93	1.86	1.06	2.13	4.26
(M2)	0.65	1.31	2.62	3.18	6.37	12.73

Here, a pair such as (3, 0.25) denotes the values of $s_0 = 3$, $b = 0.25$ (where b is the value of the active regression coefficients).

We consider the decision-rule at significance level $\alpha = 0.05$

$$\text{reject } H_{0,j} \text{ if } P_j \leq 0.05, \quad (5.1)$$

for testing single hypotheses where P_j is as in (2.13) with plugged-in estimate $\hat{\sigma}$. The considered type I error is the average over non-active variables:

$$(p - s_0)^{-1} \sum_{j \in S_0^c} \mathbb{P}[P_j \leq 0.05] \quad (5.2)$$

and the average power is

$$s_0^{-1} \sum_{j \in S_0} \mathbb{P}[P_j \leq 0.05]. \quad (5.3)$$

For multiple testing, we consider the adjusted p-value $P_{\text{corr};j}$ from (3.1): the decision is as in (5.1) but replacing P_j by $P_{\text{corr};j}$. We report the familywise error rate (FWER) $\mathbb{P}[V_{0.05} > 0]$ and the average power as in (5.3) but the latter with using $P_{\text{corr};j}$. The results

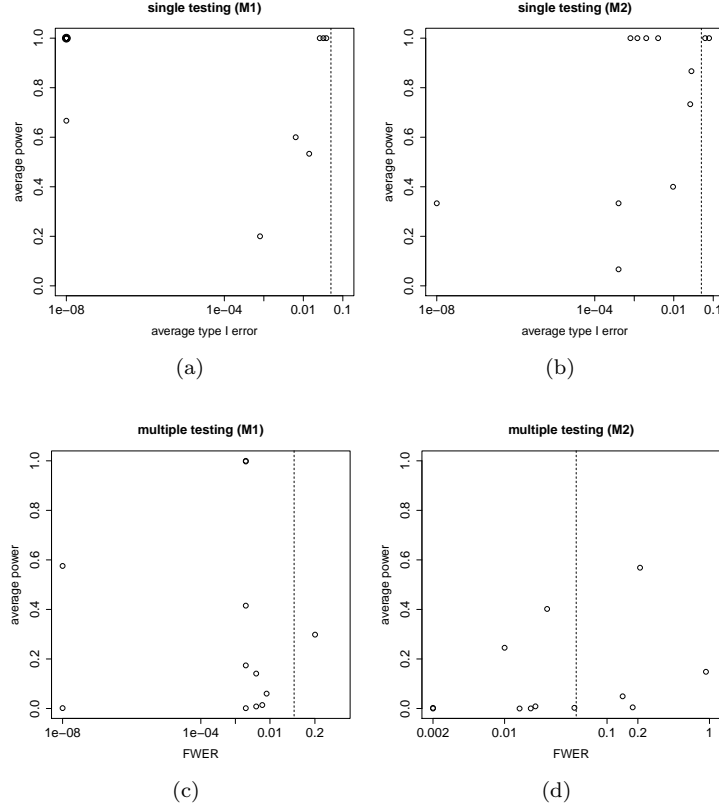


Figure 1. Simulated data as described in Section 5.1. (a) and (b): Single testing with average type I error (5.2) on x-axis (log-scale) and average power (5.3) on y-axis. (c) and (d): Multiple testing with familywise error rate on x-axis (log-scale) and average power (5.3), but using $P_{\text{corr};j}$, on y-axis. Vertical dotted line is at abscissa 0.05. Each point corresponds to a model configuration. (a) and (c): 12 model configurations generated from independent covariates (M1); (b) and (d): 12 model configurations generated from equi-dependent covariates (M2). When an error is zero, we plot it on the log-scale at abscissa 10^{-8} .

are displayed in Figure 1, based on 500 simulation runs per setting. The subfigures (c) and (d) show that the proposed method sometimes exhibits a too large familywise error rate in multiple testing. However, the corresponding number of false positives are still small except for the most extreme model configuration, see Table 3 in Supplementary Section 8.3.

5.2. Values of $P_{\mathbf{X}}$

The detection results in (4.1) and (4.2) depend on the ratio $\kappa_j = \max_{k \neq j} |(P_{\mathbf{X}})_{jk}| / |(P_{\mathbf{X}})_{jj}|$. We report in Table 1 summary statistics of $\{\kappa_j\}_j$ for various datasets. We clearly see

dataset, (n, p)	$\min_j \kappa_j$	$0.25\text{-q}\{\kappa_j\}_j$	$\text{med}\{\kappa_j\}_j$	$0.75\text{-q}\{\kappa_j\}_j$	$\max_j \kappa_j$
(M1), (100, 500)	0.21	0.27	0.29	0.31	0.44
(M1), (100, 2500)	0.27	0.34	0.36	0.39	0.54
(M2), (100, 500)	0.20	0.26	0.29	0.32	0.45
(M2), (100, 2500)	0.26	0.33	0.36	0.39	0.59
Motif, (143, 287)	0.05	0.10	0.13	0.18	0.47
Riboflavin, (71, 4088)	0.29	0.54	0.65	0.77	1.73
Leukemia, (72, 3571)	0.32	0.44	0.50	0.58	1.57
Colon, (62, 2000)	0.28	0.50	0.57	0.67	1.36
Lymphoma, (62, 4026)	0.34	0.52	0.63	0.78	1.49
Brain, (34, 5893)	0.51	0.63	0.67	0.74	2.44
Prostate, (102, 6033)	0.26	0.45	0.57	0.74	3.67
NCI, (61, 5244)	0.37	0.52	0.61	0.79	1.76

Table 1. Minimum, maximum and three quartiles of $\{\kappa_j\}_{j=1}^p$ for various designs \mathbf{X} from different datasets. The first four are from the simulation models in Section 5.1. Although not relevant for the table, “Motif” (see Section 5.3) and “Riboflavin” have a continuous response while the last six have a class label (Dettling, 2004).

that the values of κ_j are typically rather small which implies good detection properties as discussed in Section 4. Furthermore, the values $\max_{k \neq j} |(P_{\mathbf{X}})_{jk}|$ occurring in the construction of Δ_j in Section 2.4.1 are typically very small (not shown here).

5.3. Real data application

We consider a problem about motif regression for finding the binding sites in DNA sequences of the HIF1 α transcription factor. The binding sites are also called motifs, and they are typically 6-15 base pairs (with categorical values $\in \{A, C, G, T\}$) long.

The data consists of a univariate response variable Y from CHIP-chip experiments, measuring the logarithm of the binding intensity of the HIF1 α transcription factor on coarse DNA segments. Furthermore, for each DNA segment, we have abundance scores for $p = 195$ candidate motifs, based on DNA sequence data. Thus, for each DNA segment i we have $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$, where $i = 1, \dots, n_{\text{tot}} = 287$ and $p = 195$. We consider a linear model as in (1.1) and hypotheses $H_{0,j}$ for $j = 1, \dots, p = 195$: rejection of $H_{0,j}$ then corresponds to a significant motif. This dataset has been analyzed in Meinshausen et al. (2009) who found one significant motif using their p-value method for a linear model based on multiple sample splitting (which assumes the unpleasant “beta-min” condition in (1.10)).

Since the dataset has $n_{\text{tot}} > p$ observations, we take one random subsample of size $n = 143 < p = 195$. Figure 2 reports the single-testing as well as the adjusted p-values for controlling the FWER. There are two significant motifs with corresponding FWER-adjusted p-values equal to 0.009 and 0.023 (whereas the method in Meinshausen

et al. (2009) based on the total sample with n_{tot} only found one significant variable with FWER-adjusted p-value equal to 0.006; the same variable which achieves p-value 0.009 with our method). Interestingly, one of the significant motifs (with p-value 0.023) is known to be a true binding site for HIF1 α , thanks to biological validation experiments.

When compared to the Bonferroni-Holm procedure for controlling FWER based on the raw p-values as shown in Figure 2(a), we have for the variables with smallest p-values:

$$\begin{array}{ll} \text{method as in (3.1):} & 0.009, 0.023, 0.137, \\ \text{Bonferroni-Holm:} & 0.011, 0.027, 0.176. \end{array}$$

Thus, for this example, the multiple testing correction as in Section 3 does not provide large improvements in power over the Bonferroni-Holm procedure; but our method is closely related to the Westfall-Young procedure which has been shown to be asymptotically optimal for a broad class of high-dimensional problems (Meinshausen et al., 2011).

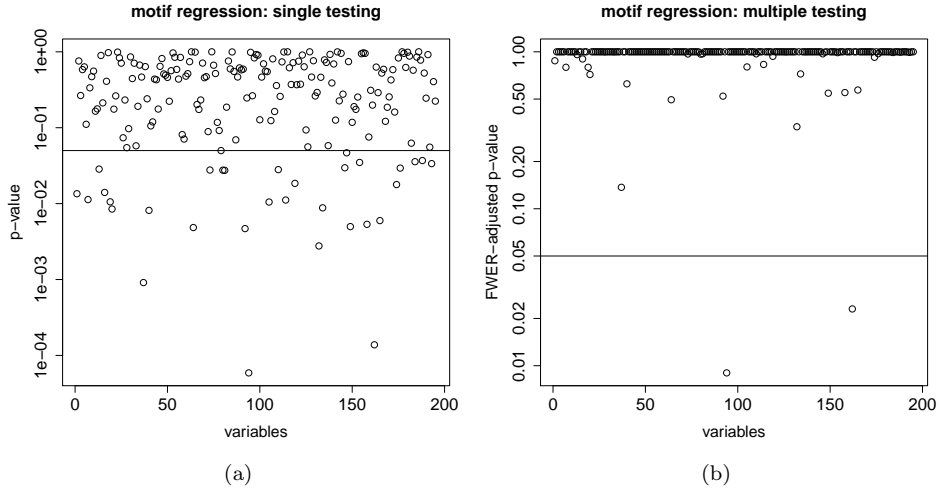


Figure 2. Motif regression with $n = 143$ and $p = 195$. (a) Single-testing p-values as in (2.13); (b) Adjusted p-values as in (3.1) for FWER control. The p-values are plotted on the log-scale. Horizontal line is at $y = 0.05$.

6. Finite sample results

We present here finite sample analogues of Theorem 1 and 2. Instead of assumption (A), we assume the following:

(A') There are constants $\Delta_j > 0$ such that

$$\mathbb{P}[\cap_{j=1}^p \{a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) \leq \Delta_j\}] \geq 1 - \kappa$$

for some (small) $0 < \kappa < 1$.

We then have the following result.

Proposition 3. Assume model (1.1) with Gaussian errors. Consider the corrected Ridge regression estimator $\hat{\beta}_{\text{corr}}$ in (2.8) with regularization parameter $\lambda > 0$, and assume (2.3) and condition (A'). Then, with probability at least $1 - \kappa$, for $j \in \{1, \dots, p\}$ and if $H_{0,j}$ holds:

$$\begin{aligned} a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr};j}| &\leq a_{n,p;j}(\sigma) |Z_j| + \Delta_j + \|a_{n,p}b(\lambda)\|_{\infty}, \\ \|a_{n,p}b(\lambda)\|_{\infty} &= \max_{j=1,\dots,p} a_{n,p;j}(\sigma) |b_j(\lambda)| \leq \frac{\lambda}{\Omega_{\min}(\lambda)^{1/2}} n^{1/2} \sigma^{-1} \|\theta^0\|_2 \lambda_{\min \neq 0}(\hat{\Sigma})^{-1}. \end{aligned}$$

Similarly, with probability at least $1 - \kappa$, for any subset $G \subseteq \{1, \dots, p\}$ and if $H_{0,G}$ holds:

$$\max_{j \in G} a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr};j}| \leq \max_{j \in G} (a_{n,p;j}(\sigma) |Z_j| + \Delta_j) + \|a_{n,p}b(\lambda)\|_{\infty}.$$

A proof is given in Supplementary Section 8.1. Due to the third statement in Lemma 1, $\Omega_{\min}(\lambda)^{-1/2}$ is bounded for a bounded range of $\lambda \in (0, C]$. Therefore, the bound for $\|a_{n,p}b(\lambda)\|_{\infty}$ can be made arbitrarily small by choosing $\lambda > 0$ sufficiently small.

Theorem 2 is a consequence of the following finite sample result.

Proposition 4. Consider the event \mathcal{E} with probability $\mathbb{P}[\mathcal{E}] \geq 1 - \kappa$ where condition (A') holds. Then, when using the corrected p-values from (3.1), with $\zeta \geq 0$ (allowing also $\zeta = 0$), we obtain approximate strong control of the familywise error rate:

$$\mathbb{P}[V_{\alpha} > 0] \leq F_Z(F_Z^{-1}(\alpha) - \zeta + 2(2\pi)^{-1/2} \|a_{n,p}b(\lambda)\|_{\infty}) + (1 - \mathbb{P}[\mathcal{E}]).$$

A proof is given in Supplementary Section 8.1. We immediately get the following bound for $\zeta \geq 0$:

$$\mathbb{P}[V_{\alpha} > 0] \leq \alpha + \sup_u |G'_Z(u)| 2(2\pi)^{-1/2} \|a_{n,p}b(\lambda)\|_{\infty} + (1 - \mathbb{P}[\mathcal{E}]).$$

7. Conclusions

We have proposed a novel construction of p-values for individual and more general hypotheses in a high-dimensional linear model with fixed design and Gaussian errors. We have restricted ourselves to max-type statistics for general hypotheses but modification to e.g. weighted sums is straightforward using the representation in Proposition 2. A key

idea is to use a linear, namely the Ridge estimator, combined with a correction for the potentially substantial bias due to the fact that the Ridge estimator is estimating the projected regression parameter vector onto the row-space of the design matrix. The fact that we can succeed with a corrected Ridge estimator in a high-dimensional context may come as a surprise, as it is well known that Ridge estimation can be very bad for say prediction: for more explanations see below. A related idea of using a linear estimator coupled with a bias correction for deriving confidence intervals has been earlier proposed by [Zhang and Zhang \(2011\)](#).

No tuning parameter. Our approach does not require the specification of a tuning parameter, except for the issue that we crudely bound the true sparsity as in (2.12); we always used $\xi = 0.05$, and the Scaled Lasso initial estimator does not require the specification of a regularization parameter. All our numerical examples were run without tuning the method to a specific setting, and error control with our p-value approach is often slightly conservative while the power seems reasonable. Regarding the latter, construction of optimal tests for $H_{0,j}$ or more general $H_{0,G}$ in a high-dimensional linear model with dependent design is an open problem. Furthermore, our method is generic which allows to test for any $H_{0,G}$ regardless whether the size of G is small or large: we present in the Supplementary Section 8.2 an additional simulation where $|G|$ is large. For multiple testing correction or for general hypotheses with sets G where $|G| > 1$, we rely on the power of simulation since analytical formulae for max-type statistics under dependence seem in-existing; yet, our simulation is extremely simple as we only need to generate dependent multivariate Gaussian random variables.

Small variance of Ridge estimator. As indicated before, it is surprising that corrected Ridge estimation performs rather well for statistical testing. Although the bias due to the projection $P_{\mathbf{X}}$ can be substantial, it is compensated by small variances $\sigma^2 n^{-1} \Omega_{jj}$ of the Ridge estimator. It is *not* true that Ω_{jj} 's become large as p increases: that is, the Ridge estimator has small variance for an individual component when p is very large, see Section 4.1. Therefore, the detection power of the method remains good as discussed in Section 4. Viewed from a different perspective, even though $|(P_{\mathbf{X}})_{jj} \beta_j^0|$ may be very small, the normalized version $a_{n,p;j}(\sigma) |(P_{\mathbf{X}})_{jj} \beta_j^0|$ can be sufficiently large for detection since $a_{n,p;j}(\sigma)$ may be very large (as the inverse of the square root of the variance). The values of $P_{\mathbf{X}}$ can be easily computed for a given problem: our analysis about sufficient conditions for detection in Section 4 could be made more complete by invoking random matrix theory for the projection $P_{\mathbf{X}}$ (assuming that \mathbf{X} is a realization of i.i.d. row-vectors whose entries are potentially dependent). However, currently, most of the results on singular values and similar quantities of \mathbf{X} are for the regime $p \leq n$ ([Vershynin, 2012](#)), which leads in our context to the trivial projection $P_{\mathbf{X}} = I$, or for the regime $p/n \rightarrow C$ with $0 \leq C < \infty$ ([El Karoui, 2008](#)).

Extensions. Obvious but partially non-trivial model extensions include random design, non-Gaussian errors or generalized linear models. From a practical point of view, the second and third issue would be most valuable. Relaxing the fixed design assumption makes part of the mathematical arguments more complicated, yet a random design is better posed in terms of identifiability.

Acknowledgments

I would like to thank Cun-Hui Zhang for fruitful discussions and Stephanie Zhang for providing an R-program for the Scaled Lasso.

References

- Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.
- Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* 37, 1705–1732.
- Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* 34, 559–583.
- Bühlmann, P., M. Kalisch, and M. Maathuis (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* 97, 261–278.
- Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.
- Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* 1, 169–194.
- Candès, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics* 35, 2313–2404.
- Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* 20, 3583–3593.
- El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics* 36, 2757–2790.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.
- Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* 70, 849–911.
- Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20, 101–148.
- Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli* 10, 971–988.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction* (Second ed.). New York: Springer.
- Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 18, 1603–1618.
- Ingster, Y., A. Tsybakov, and N. Verzelen (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* 4, 1476–1526.
- Knight, K. and W. Fu (2000). Asymptotics of Lasso-type estimators. *The Annals of Statistics* 28, 1356–1378.

- Koltchinskii, V. (2009a). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* 15, 799–828.
- Koltchinskii, V. (2009b). Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 45, 7–57.
- Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis* 52, 374–393.
- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* 34, 1436–1462.
- Meinshausen, N. and P. Bühlmann (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B* 72, 417–473.
- Meinshausen, N., M. Maathuis, and P. Bühlmann (2011). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. To appear in *Annals of Statistics*.
- Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* 104, 1671–1681.
- Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* 37, 246–270.
- Raskutti, G., M. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research* 11, 2241–2259.
- Shao, J. and X. Deng (2011). Estimation in high-dimensional linear models with deterministic design matrices. Preprint.
- Sun, T. and C.-H. Zhang (2011). Scaled sparse linear regression. arXiv:1104.4595v1.
- Tibshirani, R. (1996). Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B* 58, 267–288.
- Tropp, J. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* 50, 2231–2242.
- van de Geer, S. (2007). The deterministic Lasso. In *JSM proceedings, 2007*, 140. American Statistical Association.
- van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics* 36, 614–645.
- van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* 3, 1360–1392.
- van de Geer, S., P. Bühlmann, and S. Zhou (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics* 5, 688–749.
- Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.), *Compressed Sensing, Theory and Applications*, Chapter 5, pp. 210–268. Cambridge University Press.
- Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* 55, 2183–2202.
- Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* 104, 1512–1524.
- Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Annals of*

- Statistics* 37, 2178–2201.
- Westfall, P. and S. Young (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons.
- Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38, 894–942.
- Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics* 36, 1567–1594.
- Zhang, C.-H. and S. Zhang (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. arXiv:1110.2563v1.
- Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* 7, 2541–2563.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics* 36, 1509–1566.

8. Supplementary Section

8.1. Proofs

Proof of Proposition 1.

The statement about the bias is given in [Shao and Deng \(2011\)](#) (proof of their Theorem 1). The covariance matrix of $\hat{\beta}$ is

$$n^{-1}\Omega = n^{-1}(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}.$$

Then, for the variance we obtain $\text{Var}(\hat{\beta}_j) = n^{-1}\sigma^2\Omega_{jj} \geq n^{-1}\sigma^2\Omega_{\min}(\lambda)$. \square

Proof of Proposition 2.

We write

$$\hat{\beta}_{\text{corr};j} = (\hat{\beta}_j - \mathbb{E}[\hat{\beta}_j]) + \theta_j^0 - \sum_{k \neq j} (P_{\mathbf{X}})_{jk} \hat{\beta}_{\text{init};k} + (\mathbb{E}[\hat{\beta}_j] - \theta_j^0).$$

The result then follows by defining $Z_j = \hat{\beta}_j - \mathbb{E}[\hat{\beta}_j]$ and using that $\theta_j^0 = (P_{\mathbf{X}}\beta^0)_j = (P_{\mathbf{X}})_{jj}\beta_j^0 + \sum_{k \neq j} (P_{\mathbf{X}})_{jk}\beta_k^0$. \square

Proof of Proposition 3 (basis for proving Theorem 1).

The bound from Proposition 1 for the estimation bias of the Ridge estimator leads to:

$$\begin{aligned}
\|a_{n,p}b(\lambda)\|_\infty &= \max_{j=1,\dots,p} a_{n,p;j}(\sigma)|\mathbb{E}[\hat{\beta}_j - \theta_j^0]| \\
&\leq \frac{\lambda\|\theta^0\|_2\lambda_{\min\neq 0}(\hat{\Sigma})^{-1}}{\sigma n^{-1/2}\Omega_{jj}^{1/2}} \\
&\leq \lambda\|\theta^0\|_2\lambda_{\min\neq 0}(\hat{\Sigma})^{-1}\sigma^{-1}n^{1/2}\Omega_{\min}(\lambda)^{-1/2}.
\end{aligned}$$

By using the representation from Proposition 2, invoking assumption (A') and assuming that the null-hypothesis $H_{0,j}$ or $H_{0,G}$ holds, respectively, the proof is completed. \square

Proof of Theorem 1.

Due to the choice of $\lambda = \lambda_n$ we have that $\|a_{n,p}b(\lambda_n)\|_\infty = o(1)$ ($n \rightarrow \infty$). The proof then follows from Proposition 3 and invoking assumption (A) saying that the probabilities for the statements in Proposition 3 converge to 1 as $n \rightarrow \infty$. \square

Proof of Proposition 4 (basis for proving Theorem 2).

Consider the set \mathcal{E} where assumption (A') holds (whose probability is at least $\mathbb{P}[\mathcal{E}] \geq 1 - \kappa$). Then, on \mathcal{E} and for $j \in S_0^c$:

$$\begin{aligned}
P_j &= 2(1 - \Phi(a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j} - \Delta_j|)) \\
&\geq 2\left(1 - \Phi(a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j} - \sum_{k \neq j} (P_{\mathbf{X}})_{jk}(\hat{\beta}_{\text{init};k} - \beta_k^0)|)\right) \\
&\geq 2(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)) - 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty,
\end{aligned}$$

where in the last inequality we used Proposition 2 and Taylor's expansion. Thus, on \mathcal{E} :

$$\begin{aligned}
\min_{j \in S_0^c} P_j &\geq \min_{j \in S_0^c} 2(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)) - 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty \\
&\geq \min_{j=1,\dots,p} 2(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)) - 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}[\min_{j \in S_0^c} P_j \leq c] &\leq \mathbb{P}[\mathcal{E} \cap \{\min_{j \in S_0^c} P_j \leq c\}] + \mathbb{P}[\mathcal{E}^c] \\
&\leq \mathbb{P}[\min_{j=1,\dots,p} 2(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)) \leq c + 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty] + \mathbb{P}[\mathcal{E}^c] \\
&= F_Z(c + 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty) + \mathbb{P}[\mathcal{E}^c].
\end{aligned}$$

Using this we obtain:

$$\begin{aligned}
\mathbb{P}[V_\alpha > 0] &= \mathbb{P}[\min_{j \in S_0^c} P_{\text{corr};j} \leq \alpha] = \mathbb{P}[\min_{j \in S_0^c} P_j \leq F_Z^{-1}(\alpha) - \zeta] \\
&\leq F_Z(F_Z^{-1}(\alpha) - \zeta + 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda)\|_\infty) + \mathbb{P}[\mathcal{E}^c],
\end{aligned}$$

This completes the proof. \square

Proof of Theorem 2.

Due to the choice of $\lambda = \lambda_n$ we have that $\|a_{n,p}b(\lambda_n)\|_\infty = o(1)$ ($n \rightarrow \infty$). Furthermore, using the formulation in Proposition 4, assumption (A) translates to a sequence of sets \mathcal{E}_n with $\mathbb{P}[\mathcal{E}_n] \rightarrow 1$ ($n \rightarrow \infty$). We then use Proposition 4 and observe that for sufficiently large n : $F_Z(F_Z^{-1}(\alpha) - \zeta + 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda_n)\|_\infty) \leq F_Z(F_Z^{-1}(\alpha)) \leq \alpha$. The modification for the case with $\alpha_n \rightarrow 0$ sufficiently slowly follows analogously: note that the second last inequality in the proof above follows by monotonicity of $F_Z(\cdot)$ and $\zeta > 2(2\pi)^{-1/2}\|a_{n,p}b(\lambda_n)\|_\infty$ for n sufficiently large. This completes the proof. \square

Proof of Theorem 3.

Throughout the proof, $\alpha_n \rightarrow 0$ is converging sufficiently slowly, possibly depending on the context of the different statements we prove. Regarding statement 1: it is sufficient that for $j \in S_0$,

$$a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j}| \gg \Delta_j.$$

From Proposition 2 we see that this can be enforced by requiring

$$a_{n,p;j}(\sigma)(|(P_{\mathbf{X}})_{jj}\beta_j^0| - |\sum_{k \neq j} (P_{\mathbf{X}})_{jk}(\hat{\beta}_{\text{init};k} - \beta_k^0)| - |Z_j| - |b_j(\lambda)|) \gg \Delta_j.$$

Since $|a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk}(\hat{\beta}_{\text{init};k} - \beta_k^0)| \leq \Delta_j$, this holds if

$$|\beta_j^0| \gg \frac{1}{|(P_{\mathbf{X}})_{jj}|a_{n,p;j}(\sigma)} \max(\Delta_j, a_{n,p;j}(\sigma)Z_j, a_{n,p;j}(\sigma)b_j(\lambda)). \quad (8.1)$$

Due to the choice of $\lambda = \lambda_n$ (as in Theorem 1) we have $a_{n,p;j}(\sigma)b_j(\lambda) \leq \|a_{n,p}(\sigma b(\lambda))\|_\infty = o(1)$. Hence (8.1) holds with probability converging to one if

$$|\beta_j^0| \gg \frac{1}{|(P_{\mathbf{X}})_{jj}|a_{n,p;j}(\sigma)} \max(\Delta_j, 1),$$

completing the proof for statement 1.

For proving the second statement, we recall that

$$1 - J_G(c) = \mathbb{P}[\max_{j \in G} (a_{n,p;j}(\sigma)|Z_j| + \Delta_j) > c].$$

Denote by $W = \max_{j \in G} (a_{n,p;j}(\sigma)|Z_j| + \Delta_j) \leq \tilde{W} = \max_{j \in G} a_{n,p;j}(\sigma)|Z_j| + \max_{j \in G} \Delta_j$. Thus,

$$\mathbb{P}[W > c] \leq \mathbb{P}[\tilde{W} > c].$$

Therefore, the statement for the p-value $\mathbb{P}[P_G \leq \alpha_n]$ is implied by

$$\mathbb{P}_{\tilde{W}}[\tilde{W} > \hat{\gamma}_G] \leq \alpha_n. \quad (8.2)$$

Using the union bound and the fact that $a_{n,p;j}(\sigma)|Z_j| \sim \mathcal{N}(0,1)$ (but dependent over different values of j), we have that

$$\max_{j \in G} a_{n,p;j}(\sigma)|Z_j| = O_P(\sqrt{\log(|G|)}).$$

Therefore, (8.2) holds if

$$\hat{\gamma}_G = \max_{j \in G} a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j}| \gg \max(\max_{j \in G} \Delta_j, \sqrt{\log(|G|)}).$$

The argument is now analogous to the proof of the first statement above, using the representation from Proposition 2.

Regarding the third statement, we invoke the rough bound

$$P_{\text{corr};j} \leq pP_j,$$

with the non-truncated Bonferroni corrected p-value at the right-hand side. Hence,

$$\max_{j \in S_0} P_{\text{corr};j} \leq \alpha_n$$

is implied by

$$\max_{j \in S_0} pP_j = \max_{j \in S_0} 2p(1 - \Phi((a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j}| - \Delta_j)_+)) \leq \alpha_n.$$

Since this is a standard Gaussian two-sided tail probability, this can be enforced (for certain slowly converging α_n) by

$$\max_{j \in S_0} 2 \exp(\log(p) - (a_{n,p;j}(\sigma)|\hat{\beta}_{\text{corr};j}| - \Delta_j)_+^2/2) = o_P(1).$$

The argument is now analogous to the proof of the first statement above, using the representation from Proposition 2.

The fourth statement involves slight obvious modifications of the arguments above. \square

8.2. P-values for $H_{0,G}$ with $|G|$ large

We report here on a small simulation study for testing $H_{0,G}$ with $G = \{1, 2, \dots, 100\}$. We consider model (M2) from Section 5.1 with 4 different configurations and we use the p-value from (2.14) with corresponding decision rule for rejection of $H_{0,G}$ if the p-value is smaller or equal to the nominal level 0.05. Table 2 describes the result based on 500 independent simulations. The method works well with much better power than multiple testing of individual hypotheses but slightly worse than average power for testing individual hypotheses without multiplicity adjustment. This is largely in agreement with the theoretical results in Theorem 3. Furthermore, the type I error control is good, even for the model (M2), $p=2500, s=15, b=1$ where the multiple testing error control performs poorly.

model	$\mathbb{P}[\text{false rejection}]$	$\mathbb{P}[\text{true rejection}]$	(power mult., power indiv.)
(M2), $p = 500, s = 15, b = 0.5$	0.08	0.68	(0.14, 1.00)
(M2), $p = 500, s = 15, b = 1$	0.07	1.00	(0.58, 1.00)
(M2), $p = 2500, s = 15, b = 0.5$	0.01	0.19	(0.06, 0.40)
(M2), $p = 2500, s = 15, b = 1$	0.06	0.99	(0.30, 0.87)

Table 2. Testing of general hypothesis $H_{0,G}$ with $|G| = 100$ using the p-value in (2.14) with significance level 0.05. Second column: type I error; Third column: power; Fourth column: comparison with power using multiple testing and average power using individual testing without multiplicity adjustment (both for all p hypotheses $H_{0,j}$ ($j = 1, \dots, p$)).

8.3. Number of false positives in simulated examples

We show here the number of false positives $V = V_{0.05}$ in the simulated scenarios where the FWER was found too large. Except for model (M2), $p=2500, s=15, b=1$, the number

model	$\mathbb{P}[V = 0]$	$\mathbb{P}[V = 1]$	$\mathbb{P}[V = 2]$	$\mathbb{P}[V = 3]$	$\mathbb{P}[4 \leq V \leq 8]$	$\mathbb{P}[V \geq 9]$
(M1), $p = 2500, s = 15, b = 1$	0.798	0.186	0.016	0	0	0
(M2), $p = 500, s = 15, b = 0.5$	0.858	0.132	0.010	0	0	0
(M2), $p = 500, s = 15, b = 1$	0.790	0.186	0.020	0.004	0	0
(M2), $p = 2500, s = 15, b = 0.5$	0.822	0.170	0.008	0	0	0
(M2), $p = 2500, s = 15, b = 1$	0.076	0.222	0.268	0.200	0.234	0

Table 3. Probabilities for false positives for simulation models from Section 5.1 in scenarios where the FWER is overshooting the nominal level 0.05.

of false positives is small although the FWER is larger than 0.05. For the extreme model (M2), $p=2500, s=15, b=1$, which has a too large sparsity and a too strong signal strength, we would need to increase ξ in (2.12) to achieve better error control.

8.4. Further discussion about p-values and bounds Δ_j in assumption (A)

The p-values in (2.13) and (2.14) are crucially based on the idea of correction with the bounds Δ_j in Section 2.4.1. The essential idea is contained in Proposition 2:

$$\begin{aligned}
& a_{n,p;j}(\sigma) \hat{\beta}_{\text{corr};j} \\
&= a_{n,p;j}(\sigma) (P_{\mathbf{X}})_{jj} - a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) + a_{n,p;j}(\sigma) Z_j + \text{negligible term}.
\end{aligned}$$

There are three cases. If

$$a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) = o_P(1), \quad (8.3)$$

a correction with the bound Δ_j would not be necessary, but of course, it does not hurt in terms of type I error control. If

$$a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) \asymp V, \quad (8.4)$$

for some non-degenerate random variable V , the correction with the bound Δ_j is necessary and assuming that Δ_j is of the same order of magnitude as V , we have a balance between Δ_j and the stochastic term $a_{n,p;j}(\sigma)Z_j$. In the last case where

$$a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0) \rightarrow \infty, \quad (8.5)$$

the bound Δ_j would be the dominating element in the p-value construction. We show in Figure 3 that there is empirical evidence that (8.4) applies most often.

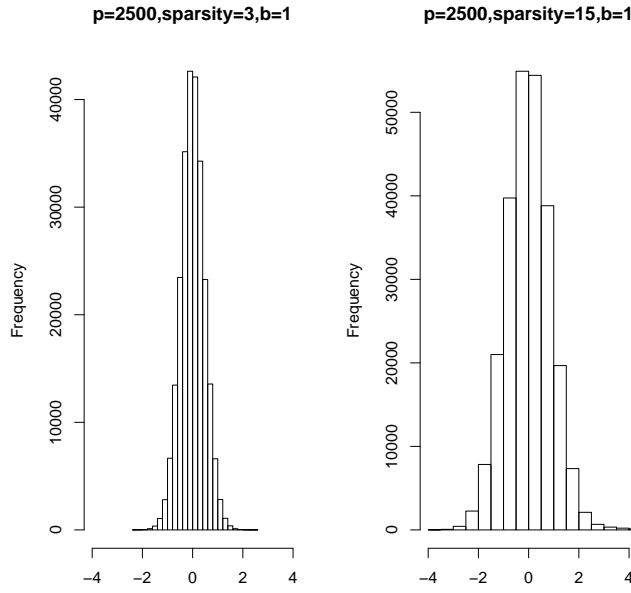


Figure 3. Histogram of projection bias $a_{n,p;j}(\sigma) \sum_{k \neq j} (P_{\mathbf{X}})_{jk} (\hat{\beta}_{\text{init};k} - \beta_k^0)$ over all values $j = 1, \dots, p$ and over 100 independent simulation runs. Left: model (M2), $p=2500, s=3, b=1$; Right: model (M2), $p=2500, s=15, b=1$.

Case (8.5) is comparable to a crude procedure which makes a hard decision about relevance of the underlying coefficients:

$$\text{if } a_{n,p;j}(\sigma) |\hat{\beta}_{\text{corr};j}| > \Delta_j \text{ holds, then } H_{0,j} \text{ is rejected,}$$

and the rejection would be “certain” corresponding to a p-value with value equal to 0; and in case of a “ \leq ” relation, the corresponding p-value would be set to one. This is an analogue to the thresholding rule:

$$\text{if } |\hat{\beta}_{\text{init};j}| > \Delta_{\text{init}} \text{ holds, then } H_{0,j} \text{ is rejected,} \quad (8.6)$$

where $\Delta_{\text{init}} \geq \|\hat{\beta}_{\text{init}} - \beta^0\|_\infty$, e.g., using a bound where $\Delta_{\text{init}} \geq \|\hat{\beta}_{\text{init}} - \beta^0\|_1$. For example, (8.6) could be the variable selection estimator with the thresholded Lasso procedure (van de Geer et al., 2011). An accurate construction of Δ_{init} for practical use is almost impossible: it depends on σ and in a complicated way on the nature of the design through e.g. the compatibility constant, see (1.7).

Our proposed bound Δ_j in (2.12) is very simple. In principle, its justification also depends on a bound for $\|\hat{\beta}_{\text{init}} - \beta^0\|_1$, but with the advantage of “robustness”. First, the bound $a_{n,p;j}(\sigma) \max_{k \neq j} |(P_{\mathbf{X}})_{jk}| \|\hat{\beta}_{\text{init}} - \beta^0\|_1$ appearing in (2.10) is not depending on σ anymore (since $\|\hat{\beta}_{\text{init}} - \beta^0\|_1$ scales linearly with σ). Secondly, the inequality in (2.10) is crude implying that Δ_j in (2.12) may still satisfy assumption (A) even if the bound of $\|\hat{\beta}_{\text{init}} - \beta^0\|_1$ is misspecified and too small. The construction of p-values as in (2.13) and (2.14) is much better for practical purposes (and for simulated examples) than using a rule as in (8.6): case (8.5) seems unlikely and our procedure enjoys “robustness” properties.